Automatic selection of indicators in a fully saturated regression

Carlos Santos · David F. Hendry · Soren Johansen

Keywords Indicators · Regression saturation · Subset selection · Model selection

Abstract We consider selecting a regression model, using a variant of the general-to-specific algorithm in PcGets, when there are more variables than observations. We look at the special case where the variables are single impulse dummies, one defined for each observation. We show that this setting is unproblematic if tackled appropriately, and obtain the asymptotic distribution of the mean and variance in a location-scale model, under the null that no impulses matter. Monte Carlo simulations confirm the null distributions and suggest extensions to highly non-normal cases.

Introduction

We consider the application of automatic general-to-specific (Gets) model selection procedures when there are more variables m than observations T in the special case that a model is saturated with a complete set of T impulse indicators, one for every observation. In this setting, the initial general unrestricted model (GUM) cannot be estimated at the outset. Instead, Hendry and Krolzig (2004) propose 'subset selection' by PcGets across combinations of candidate variables, each search path leading

C. Santos (✉)
Department of Economics and Management, Portuguese Catholic University,
Rua Diogo Botelho 1327, 4169-005 Porto, Portugal
e-mail: csantos@porto.ucp.pt

D. F. Hendry
Department of Economics, University of Oxford, Oxford, UK

S. Johansen
Department of Statistics, University of Copenhagen, Copenhagen, Denmark

to a terminal model, followed by searches across the union of these.[1] We show that their approach can be applied successfully to the selection of indicators. For general analysis of Gets, see inter alia, Hoover and Perez (1999, 2004), Krolzig and Hendry (2001), Hendry and Krolzig (2001, 2003, 2005), Campos et al. (2003), Granger and Hendry (2003), and Campos et al. (2005); details of the standard PcGets algorithm are presented in the appendix to Hendry and Krolzig (2001).

When $m > T$, all regressors cannot be entered simultaneously. Consequently, models based on combinations of subsets of $m_1 \leq T/2$ variables are explored seriatim, and a new joint model is formulated from all the terminal models thereby selected. If this union model is sufficiently small, PcGets can be applied as usual; otherwise repeated serial searches are required. Variants of this algorithm are discussed by Hendry and Krolzig (2004). Under the null that none of the T indicator variables (impulses) matters, we derive the distributions of post-selection estimators of the mean and variance in a simple location-scale data-generation-process (DGP). Monte Carlo simulations confirm the null distributions obtained.

As an analogy the PcGets search procedure attempts to sieve valuable information (regressors that genuinely matter) from 'garbage' (regressors that are in fact irrelevant, but this is not known to the investigator. Its properties, when doing so for $m << T$ are described in Hendry and Krolzig (2005). The sieving can be achieved in one step in that case, namely all candidate indicator variables are added ab initio, and checked for relevance by multi-path searches, using critical values that depend on m, T, and the investigator's perceived costs of over, versus under, selection. If the total set of candidates' exceeds the sieve's capacity, the search is conducted in stages, designed to ensure that almost all low order interactions among the regressors are examined. Here, we establish the sampling properties when $m = T + 1$ candidate variables are postulated, and interpret the outcomes. Other approaches to $m > T$ include e.g., Foster and Stine (2004).

The paper is organized as follows: Sect. 2 considers model selection when there are too many indicators for the available sample; Sect. 3 derives the mean and variance of the sampling distribution of the mean; Sect. 4 presents simulation evidence on its finite-sample accuracy, and discusses a case of a non-normal distribution; Sect. 5 concludes.

Model selection with T indicator variables

We consider the behavior for regressions which are 'saturated' by indicator variables. Let an observed random variable y_t be independently normally distributed as $y_t \sim \mathsf{IN}[\mu, \sigma_\varepsilon^2]$ for $t = 1, \ldots, T$, where $\mu \in \mathbb{R}$, $\sigma_\varepsilon^2 \in \mathbb{R}_+$ are the parameters of interest. However, an investigator is uncertain where outliers (if any) may lurk. She therefore defines a saturating set of T indicators $d_{j,t} = 1_{\{j=t\}}$, one for every j, and wishes to estimate μ and σ_ε^2 from a regression of y_t on $\{\mu, d_{j,t}, j = 1, \ldots, T-1\}$. Since a perfect fit will always result from such a regression, nothing is learned.

[1] PcGets is an Ox Package (see Doornik 2001), implementing automatic general-to-specific modeling for linear regression models, based on the theory of reduction (see, inter alia, Hendry 1995, Chapter 9).

As a first step, consider instead adding half of the indicators (e.g., $d_{j,t}$ for $j = 1, \ldots, T/2$, assuming for simplicity that T is even) together with the intercept. Thus we consider the GUM of the first step:

$$y_t = \mu + \sum_{j=1}^{T/2} \delta_j d_{j,t} + \varepsilon_t. \tag{1}$$

Hence, (1) contains $T/2$ parameters for $T/2$ impulse indicators for the first $T/2$ observations, as well as the mean and variance. Below, we consider alternative divisions of the indicators across the sample.

We find:

$$\widehat{\mu}_1 = \frac{1}{T/2} \sum_{t=T/2+1}^{T} y_t \tag{2}$$

$$s_1^2 = \frac{1}{T/2-1} \sum_{t=T/2+1}^{T} (y_t - \widehat{\mu}_1)^2 \tag{3}$$

$$\widehat{\delta}_t = y_t - \widehat{\mu}_1, \quad t = 1, \ldots, T/2 \tag{4}$$

so that:

$$\widehat{\varepsilon}_t = 0, \quad t = 1, \ldots, T/2$$
$$\widehat{\varepsilon}_t = y_t - \widehat{\mu}_1, \quad t = T/2+1, \ldots, T.$$

Because the estimators of μ and σ^2 are the usual ones for the remaining sample, we find that:

$$\mathsf{E}[\widehat{\mu}_1] = \mu \quad \text{and} \quad \mathsf{V}[\widehat{\mu}_1] = (T/2)^{-1}\sigma_\varepsilon^2$$

and

$$\mathsf{E}\big[s_1^2\big] = \sigma_\varepsilon^2.$$

Consequently, both GUM estimators are unbiased at this stage.

Next, adopting the usual PcGets approach, a parsimonious model is selected from (1) such that all mis-specification tests remain insignificant and all retained variables are significant at the desired level. That terminal model is stored, ensuring the intercept is one of the 'variables' retained by assigning it a fixed status. This selection simply involves eliminating any indicator where $|\mathsf{t}_{1,\widehat{\delta}_t}| < c_\alpha$, when the significance level c_α is used (such as that corresponding to $\alpha = 0.025$ or $\alpha = 0.01$, or, more generally, a function of T to control the false rejection rate under the null).

Now re-commence from the equivalent of (1), but entering only the other half of the impulses, namely $(\mu, d_{t,j},\ j = T/2+1, \ldots, T)$. Repeat the process to estimate μ and σ^2 by $\widehat{\mu}_2$ and s_2^2. Then, apply PcGets again, eliminating indicators where $|\mathsf{t}_{2,\widehat{\delta}_t}| < c_\alpha$

and storing the resulting parsimonious selection. Lastly, formulate a model where all significant selected indicators from the two terminal models are combined. This demonstrates that despite saturating by indicators, a feasible algorithm exists for checking every observation.

The final estimators are:

$$\widetilde{\mu} = \frac{\sum_{t=1}^{T_1} y_t 1_{\{|\mathrm{t}_{1,\widehat{\delta}_t}| < c_\alpha\}} + \sum_{t=T_1+1}^{T} y_t 1_{\{|\mathrm{t}_{2,\widehat{\delta}_t}| < c_\alpha\}}}{\sum_{t=1}^{T_1} 1_{\{|\mathrm{t}_{1,\widehat{\delta}_t}| < c_\alpha\}} + \sum_{t=T_1+1}^{T} 1_{\{|\mathrm{t}_{2,\widehat{\delta}_t}| < c_\alpha\}}} \tag{5}$$

and

$$\widetilde{\sigma}_\varepsilon^2 = \frac{\sum_{t=1}^{T_1} (y_t - \widehat{\mu}_1)^2 1_{\{|\mathrm{t}_{1,\widehat{\delta}_t}| < c_\alpha\}} + \sum_{t=T_1+1}^{T} (y_t - \widehat{\mu}_2)^2 1_{\{|\mathrm{t}_{2,\widehat{\delta}_t}| < c_\alpha\}}}{\sum_{t=1}^{T_1} 1_{\{|\mathrm{t}_{1,\widehat{\delta}_t}| < c_\alpha\}} + \sum_{t=T_1+1}^{T} 1_{\{|\mathrm{t}_{2,\widehat{\delta}_t}| < c_\alpha\}} - 1}. \tag{6}$$

The next section presents a formal analysis and derives the asymptotic properties of the estimators (5) and (6).

Although the "perfect fit" problem no longer arises, it may be thought that the huge number of $T/2$ indicators entered in each stage might induce spurious significance. However, the corresponding group of observations is simply 'dummied out' for estimating μ, which is then just the mean of the remaining sample. For an approximately normal distribution, αT outliers will occur on average under the null for a significance level α, so $\alpha T/2$ indicators will be selected on average at each stage, and αT overall: an indicator will be significant at level α if and only if there is an α-level outlier at that observation. Under the null, therefore, the proposed procedure is close to finding outliers relative to the all sample mean $\hat{\mu}$ and variance $\hat{\sigma}^2$.

Additional regressors will entail an inability to add half the indicators at each stage, and may necessitate exploring many combinations, but do not, otherwise, affect the analysis.

Conversely, testing many different forms of hypothesis, could alter the null rejection frequency. For example, checking the joint significance of all possible pairs, triplets, etc. will not deliver a null rejection frequency α. This is not a serious issue under the null hypothesis that only $\delta_i = 0$ for all i; but researcher may have the temptation to consider (e.g.,) step shifts where blocks of δ_i take the same values. To control the null rejection frequency, the number of classes of hypotheses has to be controlled, and one way of achieving that goal is to restrict such hypothesis searches to situations where the null has been rejected. Conditional on that occurrence, than many alternatives of how to form an index of the retained indicators can be entertained, which will not affect the null rejection frequency: Hendry and Santos (2005) show that after selecting indicators, indexes thereof can be formed without distorting inference.

There is a selection effect on the mean and variance estimates in the final model, similar to 'trimming', and the approximate distributions are derived in Sect. 3. The three-stage PcGets procedure is difficult to analyze directly, so the approach therein is to eliminate half of the sample by adding half the indicators (see Salkever 1976), then select outliers in the remaining half. Next, the converse half-sample is removed and the other group of outliers detected. This procedure entails that, on both steps, outliers in the saturated half are also removed, so is close to the third stage of PcGets.

The analysis then derives the distribution of the mean based on the two subsample means, as well as the mean of the error variance. In fact, since an exact sample split is not needed, and may sometimes be undesirable, the analysis allows for a general split, and Sects. 3.3 and 3.4 consider the possibility that many splits are used.

The role of the Monte Carlo experiments in Sect. 4 is, therefore, to check that the theory is indeed closely relevant to the PcGets procedure in small samples when the null distribution is a standard normal, as well as being relevant for other distributions.

Sampling distributions

We first derive the sampling distribution of $\tilde{\mu}$ under the null after dummy saturation, then consider the impact of saturation on $\tilde{\sigma}_\varepsilon^2$.

3.1 Asymptotic distribution of $\tilde{\mu}$

We derive the asymptotic distribution of $\widetilde{\mu}$ calculated under the assumptions that the first analysis has T_1 dummies and the second has $T_2 = T - T_1$ dummies, whereas the data generating process has IID variables.

Theorem 1 *Let $y_1, \ldots, y_T$ be IID with a symmetric continuous density $f(\cdot)$ with mean μ and $\mathsf{E}[y_t^8] < \infty$. Let $T = T_1 + T_2$, and assume that $T_1/T \to \lambda_1$ and $T_2/T \to \lambda_2$ where $0 < \lambda_1, \lambda_2 < 1$, with $\lambda_1 + \lambda_2 = 1$. Then the limit distribution of the estimator $\widetilde{\mu}$, see (5), is given by*:

$$T^{1/2}(\widetilde{\mu} - \mu) \xrightarrow{D} \mathsf{N}\left[0, \sigma_\varepsilon^2 \sigma_\mu^2\right], \tag{7}$$

where

$$\sigma_\mu^2 = \left(\int_{-c_\alpha}^{c_\alpha} f(\varepsilon) d\varepsilon \right)^{-2} \left[\int_{-c_\alpha}^{c_\alpha} \varepsilon^2 f(\varepsilon) d\varepsilon (1 + 4c_\alpha f(c_\alpha)) + \left(\frac{\lambda_1^2}{\lambda_2} + \frac{\lambda_2^2}{\lambda_1} \right) (2c_\alpha f(c_\alpha))^2 \right].$$

Note that $\int_{-c_\alpha}^{c_\alpha} f(\varepsilon) d\varepsilon = 1 - \alpha$, and for the normal distribution, $f(\varepsilon) = \frac{1}{\sigma_\varepsilon} \phi(\frac{\varepsilon}{\sigma_\varepsilon})$, we find the expression:

$$\int_{-c_\alpha}^{c_\alpha} \varepsilon^2 \phi(\varepsilon) d\varepsilon = \int_{-c_\alpha}^{c_\alpha} \phi(\varepsilon) d\varepsilon - 2c_\alpha \phi(c_\alpha)$$

so that under normality for an equal split ($\lambda_1 = \lambda_2$):

$$\sigma_\mu^2 = \frac{1}{(1-\alpha)} \left(1 + 4c_\alpha \phi(c_\alpha) - \frac{2c_\alpha \phi(c_\alpha)}{(1-\alpha)} \left[1 + 2c_\alpha \phi(c_\alpha) \right] \right). \tag{8}$$

Proof There is no loss of generality in setting $\sigma_\varepsilon^2 = 1$, and we let $c = c_\alpha$. The estimator satisfies:

$$
\begin{aligned}
&T^{1/2}(\widetilde{\mu} - \mu) \\
&= \frac{T^{-1/2}\left(\sum_{t=1}^{T_1} \varepsilon_t 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_1| \le c s_1 \sqrt{1+T_2^{-1}}\right\}} + \sum_{t=T_1+1}^{T} \varepsilon_t 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_2| \le c s_2 \sqrt{1+T_1^{-1}}\right\}}\right)}{T^{-1}\left(\sum_{t=1}^{T_1} 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_1| \le c s_1 \sqrt{1+T_2^{-1}}\right\}} + \sum_{t=T_1+1}^{T} 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_2| \le c s_2 \sqrt{1+T_1^{-1}}\right\}}\right)} \\
&= \frac{B_T}{M_T}.
\end{aligned}
$$

We show that B_T converges in distribution to a normal distribution, and M_T converges in probability to a constant. The problem is the dependence structure due to the appearance of $(\bar{\varepsilon}_1, s_1^2)$ and $(\bar{\varepsilon}_2, s_2^2)$ in the selection variables. We therefore define the simpler variables which are sums of IID variables:

$$
\begin{aligned}
K_T &= T^{-1}\left(\sum_{t=1}^{T_1} 1_{\{|\varepsilon_t| \le c\}} + \sum_{t=T_1+1}^{T} 1_{\{|\varepsilon_t| \le c\}}\right), \\
C_T &= T^{-1/2}\left(\sum_{t=1}^{T_1} \left(\varepsilon_t 1_{\{|\varepsilon_t| \le c\}} + 2cf(c)\bar{\varepsilon}_1\right) + \sum_{t=T_1+1}^{T} \left(\varepsilon_t 1_{\{|\varepsilon_t| \le c\}} + 2cf(c)\bar{\varepsilon}_2\right)\right).
\end{aligned}
$$

We want to approximate B_T/M_T by C_T/K_T and so write:

$$
T^{1/2}(\widetilde{\mu} - \mu) = \frac{B_T}{M_T} = \frac{(B_T - C_T) + C_T}{(M_T - K_T) + K_T}.
$$

From the law of large numbers:

$$
K_T \xrightarrow{\mathsf{P}} \int_{-c}^{c} f(\varepsilon) d\varepsilon. \tag{9}
$$

By symmetry of the distribution, $\mathsf{E}[C_T] = 0$. Furthermore,

$$
C_T = T^{-1/2}\left(\sum_{t=1}^{T_1}\left(\varepsilon_t 1_{\{|\varepsilon_t| \le c\}} + \frac{\lambda_2}{\lambda_1} 2cf(c)\varepsilon_t\right) + \sum_{t=T_1+1}^{T}\left(\varepsilon_t 1_{\{|\varepsilon_t| \le c\}} + \frac{\lambda_1}{\lambda_2} 2cf(c)\varepsilon_t\right)\right).
$$

So, from the central limit theorem, C_T is asymptotically normal with mean zero and variance:

$$\lambda_1\left[\mathsf{E}\left[\varepsilon_t^2 1_{\{|\varepsilon_t|\le c\}}\right]+\left(\frac{\lambda_2}{\lambda_1}\right)^2(2cf(c))^2+4cf(c)\frac{\lambda_2}{\lambda_1}\mathsf{E}\left[\varepsilon_t^2 1_{\{|\varepsilon_t|\le c\}}\right]\right]$$
$$+\lambda_2\left[\mathsf{E}\left[\varepsilon_t^2 1_{\{|\varepsilon_t|\le c\}}\right]+\left(\frac{\lambda_1}{\lambda_2}\right)^2(2cf(c))^2+4cf(c)\frac{\lambda_1}{\lambda_2}\mathsf{E}\left[\varepsilon_t^2 1_{\{|\varepsilon_t|\le c\}}\right]\right]$$
$$=\mathsf{E}\left[\varepsilon_t^2 1_{\{|\varepsilon_t|\le c\}}\right](1+4cf(c))+\left(\frac{\lambda_2^2}{\lambda_1}+\frac{\lambda_1^2}{\lambda_2}\right)(2cf(c))^2$$

which together with (9) gives the expression for σ_μ^2. We therefore only have to prove that:

$$M_T-K_T\xrightarrow{\mathsf{P}}0, \tag{10}$$
$$B_T-C_T\xrightarrow{\mathsf{P}}0. \tag{11}$$

To prove (10) we note that it is enough to show that:

$$D_T=T_1^{-1}\sum_{t=1}^{T_1}\left(1_{\left\{|\varepsilon_t-\bar{\varepsilon}_1|\le cs_1\sqrt{1+T_2^{-1}}\right\}}-1_{\{|\varepsilon_t|\le c\}}\right)\xrightarrow{\mathsf{P}}0 \tag{12}$$

since the other one follows by replacing subscript 1 by 2. Let $u=\bar{\varepsilon}_1$ and $v=c(s_1\sqrt{1+T_2^{-1}}-1)$ and apply the inequality:

$$\left|1_{\left\{|\varepsilon_t-\bar{\varepsilon}_1|\le cs_1\sqrt{1+T_2^{-1}}\right\}}-1_{\{|\varepsilon_t\le c\}}\right|=|1_{\{|\varepsilon_t-u|\le c+v\}}-1_{\{|\varepsilon_t|\le c\}}|$$
$$\le 1_{\{|\varepsilon_t-c|\le|u|+|v|\}}+1_{\{|\varepsilon_t+c|\le|u|+|v|\}} \tag{13}$$

to find:

$$T_1^{-1}\mathsf{E}_{uv}|D_T|\le\int_{c-|u|-|v|}^{c+|u|+|v|}\varepsilon f(\varepsilon)d\varepsilon+\int_{-c-|u|-|v|}^{-c+|u|+|v|}\varepsilon f(\varepsilon)d\varepsilon=h(|u|+|v|)$$

which is bounded and continuous in $|u|+|v|$ by the assumptions. Because $|u|+|v|\xrightarrow{\mathsf{P}}0$, we then get, by taking expectations, that:

$$T_1^{-1}\mathsf{E}|D_T|\le\mathsf{E}\left[h\left(|u|+|v|\right)\right]\to h(0)=0.$$

This shows that $D_T \xrightarrow{\mathsf{P}} 0$ and hence (10). We next prove (11). It is enough to show that:

$$R_T = T_1^{-1/2} \sum_{t=1}^{T_1} \left(\varepsilon_t 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_1| \leq c s_1 \sqrt{1+T_2^{-1}}\right\}} - \varepsilon_t 1_{\{|\varepsilon_t| \leq c\}} - 2cf(c)\bar{\varepsilon}_1 \right) \xrightarrow{\mathsf{P}} 0.$$

By symmetry, we have that $\mathsf{E}[R_T] = 0$, and we want to show that $\mathsf{V}[R_T] \to 0$. To find the variance, we again condition on $\bar{\varepsilon}_1 = u$ and $c(s_1\sqrt{1+T_2^{-1}} - 1) = v$, which are independent of the variables $\varepsilon_1, \ldots, \varepsilon_{T_1}$, which remain IID, and find:

$$\begin{aligned}
\mathsf{E}_{uv}[R_T] &= T_1^{1/2} \mathsf{E}\left[\varepsilon_t 1_{\{|\varepsilon_t - u| \leq c+v\}} - \varepsilon_t 1_{\{|\varepsilon_t| \leq c\}} - 2cf(c)u\right] \\
&= T_1^{1/2} \left(\int_{-c-v+u}^{c+v+u} \varepsilon f(\varepsilon) d\varepsilon - \int_{-c}^{c} \varepsilon f(\varepsilon) d\varepsilon - 2cf(c)u \right).
\end{aligned}$$

From Taylor's formulae with remainder term, we find for a differentiable function:

$$g(c+h) = g(c) + hg(c^*) = g(c) + hg(c) + h(g(c^*) - g(c)), \quad |c - c^*| \leq |h|.$$

This implies that, using $f(c) = f(-c)$:

$$\begin{aligned}
\int_{-\infty}^{c+v+u} \varepsilon f(\varepsilon) d\varepsilon &= \int_{-\infty}^{c} \varepsilon f(\varepsilon) d\varepsilon + (u+v)cf(c) + (u+v)\left(c^* f(c^*) - cf(c)\right), \\
\int_{-\infty}^{-c-v+u} \varepsilon f(\varepsilon) d\varepsilon &= \int_{-\infty}^{-c} \varepsilon f(\varepsilon) d\varepsilon - (u-v)cf(c) + (u-v)(-c^{**} f(c^{**}) + cf(c)).
\end{aligned}$$

Subtracting these expressions, we find that:

$$|\mathsf{E}_{uv}[R_T]| \leq T_1^{1/2}(|u| + |v|)(|c^* f(c^*) - cf(c)| + |c^{**} f(c^{**}) - cf(c)|).$$

Hence:

$$\begin{aligned}
\mathsf{V}\left[\mathsf{E}_{uv}[R_T]\right] &\leq \mathsf{E}\left[\mathsf{E}_{uv}[R_T]\right]^2 \\
&\leq T_1 \mathsf{E}\left[|u| + |v|\right]^2 \left(|c^* f(c^*) - cf(c)| + |c^{**} f(c^{**}) - cf(c)|\right)^2 \\
&\leq 2T_1 \mathsf{E}\left[u^2 + v^2\right] \left(|c^* f(c^*) - cf(c)| + |c^{**} f(c^{**}) - cf(c)|\right)^2 \\
&\leq 2^{3/2} T_1 \left(\mathsf{E}\left[u^4 + v^4\right]\right)^{1/2} \\
&\quad \times \mathsf{E}\left[\left(|c^* f(c^*) - cf(c)| + |c^{**} f(c^{**}) - cf(c)|\right)^4\right]^{1/2},
\end{aligned}$$

where we used the inequality $(a+b)^2 \le 2(a^2+b^2)$ twice and the Cauchy–Schwartz inequality to separate the expectations. Note that because $|\varepsilon_t|$ has a finite mean, we have $|c|f(c) \to 0$, $|c| \to \infty$, so that the continuity of $f(\cdot)$ implies $|c|f(c)$ is a bounded continuous function. Because

$$\max(|c-c^{**}|, |c-c^{*}|) \le |u|+|v| = |\bar{\varepsilon}_1| + c\left|s_1\sqrt{1+T_2^{-1}} - 1\right| \xrightarrow{\mathsf{P}} 0$$

it follows that $c^* \xrightarrow{\mathsf{P}} c$ and $c^{**} \xrightarrow{\mathsf{P}} c$, so that:

$$\mathsf{E}\left((|c^* f(c^*) - cf(c)| + |c^{**} f(c^{**}) - cf(c)|)^4\right) c^* \to 0.$$

We then have to prove that $T_1^2\mathsf{E}[u^4+v^4]$ is bounded. The first term is

$$T_1^2\mathsf{E}\left[\bar{\varepsilon}_1^4\right] = T_1^{-1}\mathsf{E}\left[\varepsilon_1^4\right] + 3(1-T_1^{-1})$$

using that $\mathsf{E}[\varepsilon_1] = \mathsf{E}[\varepsilon_1^3] = 0$ and $\mathsf{E}[\varepsilon_1^2] = 1$. This is bounded when we assume finite fourth moment. Next,

$$T_1^2\mathsf{E}\left[s_1\sqrt{1+T_2^{-1}} - 1\right]^4 \le 8\left[T_1^2\mathsf{E}[s_1-1]^4\left(1+T_2^{-1}\right)^2 + T_1^2\left(1-\sqrt{1+T_2^{-1}}\right)^4\right].$$

The factor $(1+T_2^{-1})^2$ and the term $T_1^2(1-\sqrt{1+T_2^{-1}})^4$ are bounded, and we evaluate:

$$\begin{aligned} T_1^2\mathsf{E}[s_1-1]^4 &\le T_1^2\mathsf{E}[s_1^2-1]^4 \\ &= T_1^{-1}\mathsf{E}[\varepsilon_t^2-1]^4 + 3(1-T_1^{-1})\left(\mathsf{E}\left[\varepsilon_t^2-1\right]^2\right)^2 \end{aligned}$$

which is bounded when ε_t has moments of order eight. Thus the first factor $T_1^2\mathsf{E}(u^4+v^4)$ is bounded and therefore:

$$\mathsf{V}\left[\mathsf{E}_{uv}\left[R_T\right]\right] \to 0. \tag{14}$$

Next, we consider $\mathsf{E}[\mathsf{V}_{uv}[R_T]]$ and, find using inequality (13), that:

$$\begin{aligned} \mathsf{V}_{uv}[R_T] &= \mathsf{E}\left[\varepsilon_t 1_{\{|\varepsilon_t - u| \le c+v\}} - \varepsilon_t 1_{\{|\varepsilon_t| \le c\}}\right]^2 \qquad (15) \\ &\le \int_{-c-|u|-|v|}^{-c+|u|+|v|} \varepsilon^2 f(\varepsilon)d\varepsilon + \int_{c-|u|-|v|}^{c+|u|+|v|} \varepsilon^2 f(\varepsilon)d\varepsilon \end{aligned}$$

which is a bounded continuous function of $|u| + |v|$, so that:

$$\mathsf{E}[\mathsf{V}_{uv}[R_T]] \to 0. \tag{16}$$

Combining (14) and (16) we see that $\mathsf{V}[R_T] \to 0$, which completes the proof of (11). □

3.2 The probability limit of $\tilde{\sigma}_\varepsilon^2$

Theorem 2 *Under the assumptions of Theorem* 1 *it holds that the estimator* $\widetilde{\sigma}_\varepsilon^2$, *see* (6), *has the limit*:

$$\widetilde{\sigma}_\varepsilon^2 \xrightarrow{\mathsf{P}} \frac{\int_{-c_\alpha}^{c_\alpha} \varepsilon^2 f(\varepsilon)d\varepsilon}{\int_{-c_\alpha}^{c_\alpha} f(\varepsilon)d\varepsilon} = \mathsf{V}[\varepsilon||\varepsilon| < c_\alpha].$$

For the normal distribution, $f(\varepsilon) = \frac{1}{\sigma_\varepsilon}\phi(\frac{\varepsilon}{\sigma_\varepsilon})$, we find the expression:

$$\frac{\int_{-c_\alpha}^{c_\alpha} \varepsilon^2 \phi(\varepsilon)d\varepsilon}{\int_{-c_\alpha}^{c_\alpha} \phi(\varepsilon)d\varepsilon} = \sigma_\varepsilon^2\left(1 - \frac{2c_\alpha\phi(c_\alpha)}{1-\alpha}\right).$$

Proof The technique is the same as in the proof of Theorem 1. We let $\sigma_\varepsilon^2 = 1$, and let $c = c_\alpha$. We first note that, see (6), $\widetilde{\sigma}_\varepsilon^2 = \frac{D_T}{L_T} + H_T$, where:

$$\frac{D_T}{L_T} = \frac{T^{-1}\sum_{t=1}^{T_1} \varepsilon_t^2 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_1| \le cs_1\sqrt{1+T_2^{-1}}\right\}} + \sum_{t=T_1+1}^{T} \varepsilon_t^2 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_2| \le cs_2\sqrt{1+T_1^{-1}}\right\}}}{T^{-1}\sum_{t=1}^{T_1} 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_1| \le cs_1\sqrt{1+T_2^{-1}}\right\}} + \sum_{t=T_1+1}^{T} 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_2| \le cs_2\sqrt{1+T_1^{-1}}\right\}}},$$

$$H_T = \frac{(\mu - \widehat{\mu}_1)^2 \sum_{t=1}^{T_1} 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_1| \le cs_1\sqrt{1+T_2^{-1}}\right\}} + (\mu - \widehat{\mu}_2)^2 \sum_{t=T_1+1}^{T} 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_2| \le cs_2\sqrt{1+T_1^{-1}}\right\}}}{\sum_{t=1}^{T_1} 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_1| \le cs_1\sqrt{1+T_2^{-1}}\right\}} + \sum_{t=T_1+1}^{T} 1_{\left\{|\varepsilon_t - \bar{\varepsilon}_2| \le cs_2\sqrt{1+T_1^{-1}}\right\}}}.$$

The last term, H_T, tends to zero in probability because $\widehat{\mu}_1 \xrightarrow{\mathsf{P}} \mu$ and $\widehat{\mu}_2 \xrightarrow{\mathsf{P}} \mu$. From (9), we know that $K_T \xrightarrow{\mathsf{P}} \int_{-c}^{c} f(\varepsilon)d\varepsilon$. We define the sum of independent variables and apply the law of large numbers to find:

$$E_T = T^{-1}\left(\sum_{t=1}^{T_1} \varepsilon_t^2 1_{\{|\varepsilon_t| \le c\}} + \sum_{t=T_1+1}^{T} \varepsilon_t^2 1_{\{|\varepsilon_t| \le c\}}\right) \xrightarrow{\mathsf{P}} \int_{-c}^{c} \varepsilon^2 f(\varepsilon)d\varepsilon.$$

We next have to show that $E_T - D_T \xrightarrow{\mathsf{P}} 0$. It is clearly enough to prove that:

$$T_1^{-1}\sum_{t=1}^{T_1} \varepsilon_t^2 \left(1_{\left\{|\varepsilon_t - \bar{\varepsilon}_1| \le cs_1\sqrt{1+T_2^{-1}}\right\}} - 1_{\{|\varepsilon_t| \le c\}}\right) \xrightarrow{\mathsf{P}} 0.$$

Conditioning on u and v, we find using (13) that:

$$\begin{aligned}
&\mathsf{E}_{uv}\left|T_1^{-1}\sum_{t=1}^{T_1}\varepsilon_t^2\left(1_{\{|\varepsilon_t-\bar{\varepsilon}_1|\le cs_1\sqrt{1+T_2^{-1}}\}}-1_{\{|\varepsilon_t\le c\}}\right)\right| \\
&\le \mathsf{E}\left[\varepsilon_t^2\left(1_{\{|\varepsilon_t-c|\le|u|+|v|\}}+1_{\{|\varepsilon_t+c|\le|u|+|v|\}}\right)\right] \\
&\le \int_{c-|u|-|v|}^{c+|u|+|v|}\varepsilon^2 f(\varepsilon)d\varepsilon+\int_{-c-|u|-|v|}^{-c+|u|+|v|}\varepsilon^2 f(\varepsilon)d\varepsilon
\end{aligned}$$

[see (15)]. This is a bounded and continuous function of $|u|+|v|$ and hence the expectation tends to zero. □

3.3 Many splits

We split the data into $I_j,\ j=1,\ldots,m$ with $T_j=\lambda_j T$ elements and estimators $\bar{y}_j, s_j^2$ and define

$$\begin{aligned}
T_{-j} &= \sum_{k\neq j}T_k = T-T_j,\quad \lambda_{-j}=1-\lambda_j \\
\bar{y}_{-j} &= \frac{\sum_{t\notin I_j}y_t}{\sum_{t\notin I_j}1}=\frac{\sum_{k\neq j}T_k\bar{y}_k}{\sum_{k\neq j}T_k} \\
s_{-j}^2 &= \frac{\sum_{k\neq j}(T_k-1)s_k^2}{\sum_{k\neq j}(T_k-1)} \\
\widetilde{\mu} &= \frac{\sum_{j=1}^m\sum_{t\in I_j}y_t 1_{\left\{|y_t-\bar{y}_{-j}|<c_\alpha s_{-j}\sqrt{1+T_{-j}^{-1}}\right\}}}{\sum_{j=1}^m\sum_{t\in I_j}1_{\left\{|y_t-\bar{y}_{-j}|<c_\alpha s_{-j}\sqrt{1+T_{-j}^{-1}}\right\}}}
\end{aligned}\tag{17}$$

and

$$\widetilde{\sigma}_\varepsilon^2=\frac{\sum_{j=1}^m\sum_{t\in I_j}(y_t-\bar{y}_{-j})^2 1_{\left\{|y_t-\bar{y}_{-j}|<c_\alpha s_{-j}\sqrt{1+T_{-j}^{-1}}\right\}}}{\sum_{j=1}^m\sum_{t\in I_j}1_{\left\{|y_t-\bar{y}_{-j}|<c_\alpha s_{-j}\sqrt{1+T_{-j}^{-1}}\right\}}}.\tag{18}$$

3.4 Asymptotic distributions of $\widetilde{\mu}$ and limit of $\widetilde{\sigma}_\varepsilon^2$

Theorem 3 *Let $y_1,\ldots,y_T$ be IID with a symmetric continuous density $f(\cdot)$ with mean μ and $\mathsf{E}[y_t^8]<\infty$. Let $T=\sum_{j=1}^m T_j$, and assume that $T_j/T\to\lambda_j$, where $0<\lambda_j<1$, with $\sum_{j=1}^m\lambda_j=1$. Then the limit distribution of the estimator $\widetilde{\mu}$, see (17), is given by*:

$$T^{1/2}(\widetilde{\mu}-\mu)\xrightarrow{D}\mathsf{N}\left[0,\sigma_\varepsilon^2\sigma_\mu^2\right],\tag{19}$$

where

$$\sigma_\mu^2 = \left(\int_{-c_\alpha}^{c_\alpha} f(\varepsilon) d\varepsilon \right)^{-2} \times \left[\int_{-c_\alpha}^{c_\alpha} \varepsilon^2 f(\varepsilon) d\varepsilon (1 + 4c_\alpha f(c_\alpha)) + \sum_{j=1}^{m} \lambda_j \left(\sum_{k \neq j} \frac{\lambda_k}{1 - \lambda_k} \right)^2 (2c_\alpha f(c_\alpha))^2 \right].$$

If in particular $T_1 = \cdots = T_m$, *then* $\sum_{j=1}^m \lambda_j (\sum_{k \neq j} \frac{\lambda_k}{\lambda_{-k}})^2 = 1$.

Proof There is no loss of generality in setting $\sigma_\varepsilon^2 = 1$, and we let $c = c_\alpha$. The estimator satisfies:

$$T^{1/2}(\widetilde{\mu} - \mu) = \frac{T^{-1/2} \sum_{j=1}^m \sum_{t \in I_j} \varepsilon_t 1_{\left\{ |\varepsilon_t - \bar{\varepsilon}_{-j}| < c s_{-j} \sqrt{1 + T_{-j}^{-1}} \right\}}}{T^{-1} \sum_{j=1}^m \sum_{t \in I_j} 1_{\left\{ |\varepsilon_t - \bar{\varepsilon}_{-j}| < c s_{-j} \sqrt{1 + T_{-j}^{-1}} \right\}}} = \frac{S_T}{W_T}.$$

We show that S_T converges in distribution to a normal distribution, and W_T converges in probability to a constant. The problem is the dependence structure due to the appearance of $(\bar{\varepsilon}_{-j}, s_{-j}^2)$ in the selection variables. We therefore define the simpler variables which are sums of IID variables:

$$Q_T = T^{-1} \sum_{j=1}^m \sum_{t \in I_j} 1_{\{|\varepsilon_t| < c\}},$$

$$U_T = T^{-1/2} \sum_{j=1}^m \sum_{t \in I_j} (\varepsilon_t 1_{\{|\varepsilon_t| \leq c\}} + 2cf(c)\bar{\varepsilon}_{-j}).$$

We want to approximate S_T / W_T by Q_T / U_T and so write:

$$T^{1/2}(\widetilde{\mu} - \mu) = \frac{S_T}{W_T} = \frac{(S_T - U_T) + U_T}{(W_T - Q_T) + Q_T}.$$

From the law of large numbers:

$$Q_T \xrightarrow{\mathsf{P}} \int_{-c}^{c} f(\varepsilon) d\varepsilon. \tag{20}$$

By symmetry of the distribution, $\mathsf{E}[U_T] = 0$. Furthermore,

$$U_T = T^{-1/2} \sum_{j=1}^{m} \sum_{t \in I_j} (\varepsilon_t 1_{\{|\varepsilon_t| \le c\}} + 2cf(c)\bar{\varepsilon}_{-j})$$

$$= T^{-1/2} \sum_{j=1}^{m} \sum_{t \in I_j} \left(\varepsilon_t 1_{\{|\varepsilon_t| \le c\}} + 2cf(c)\varepsilon_t \left[\sum_{k \neq j} \frac{T_k}{T_{-k}} \right] \right).$$

So, from the central limit theorem, U_T is asymptotically normal with mean zero and variance:

$$T^{-1} \left[\sum_{j=1}^{m} T_j \left(\mathsf{E}\left[\varepsilon_t^2 1_{\{|\varepsilon_t| \le c\}}\right] + \left[\sum_{k \neq j} \frac{T_k}{T_{-k}} \right]^2 [2cf(c)]^2 + 4cf(c) \left[\sum_{k \neq j} \frac{T_k}{T_{-k}} \right] \mathsf{E}\left[\varepsilon_t^2 1_{\{|\varepsilon_t| \le c\}}\right] \right) \right]$$

$$= \mathsf{E}\left[\varepsilon_t^2 1_{\{|\varepsilon_t| \le c\}}\right] \left(1 + 4cf(c)T^{-1} \sum_{j=1}^{m} T_j \sum_{k \neq j} \frac{T_k}{T_{-k}} \right) + T^{-1} \sum_{j=1}^{m} T_j \left(\sum_{k \neq j} \frac{T_k}{T_{-k}} \right)^2 (2cf(c))^2$$

$$= \mathsf{E}\left[\varepsilon_t^2 1_{\{|\varepsilon_t| \le c\}}\right] \left(1 + 4cf(c) \sum_{j=1}^{m} \lambda_j \sum_{k \neq j} \frac{\lambda_k}{\lambda_{-k}} \right) + \sum_{j=1}^{m} \lambda_j \left(\sum_{k \neq j} \frac{\lambda_k}{\lambda_{-k}} \right)^2 (2cf(c))^2.$$

Next we show that

$$\sum_{j=1}^{m} \lambda_j \sum_{k \neq j} \frac{\lambda_k}{\lambda_{-k}} = \sum_{k \neq j} \frac{\lambda_j \lambda_k}{1 - \lambda_k} = \sum_{k=1}^{m} \sum_{j \neq k} \frac{\lambda_j \lambda_k}{1 - \lambda_k} = \sum_{k=1}^{m} \frac{(1 - \lambda_k)\lambda_k}{1 - \lambda_k} = 1. \tag{21}$$

which together with (9) gives is the expression for σ_μ^2. If in particular $\lambda_t = m^{-1}$, then

$$\sum_{j=1}^{m} \lambda_j \left(\sum_{k \neq j} \frac{\lambda_k}{\lambda_{-k}} \right)^2 = \sum_{j=1}^{m} \frac{1}{m} \left(\sum_{k \neq j} \frac{\frac{1}{m}}{1 - \frac{1}{m}} \right)^2 = \sum_{j=1}^{m} \frac{1}{m} \left[(m-1) \frac{1}{m-1} \right]^2 = 1.$$

□

Monte Carlo experiments

We first examine the properties of the retained impulses under normality, checking that the selection delivers retention rates which match the binomial expansion of $(\alpha + [1 - \alpha])^T$ despite the sequential selection. Next, we check that different sample splits ($T/3$ etc.) do not affect the null outcome. Then we investigate the empirical distributions of $\widetilde{\sigma}_\varepsilon^2$ and $\widetilde{\mu}$, under the null, to check the small-sample relevance of the derivations in Sect. 3. We also briefly consider the impact of saturation in a highly non-normal case, namely a $\mathsf{t}_{(4)}$-distributed error.

Table 1 Null rejection frequencies of impulse indicators when the DGP is (22); $T = 100$

Mean $RF_{\alpha=0.05}$	Mean $RF_{\alpha=0.025}$	Mean $RF_{\alpha=0.01}$
0.0499	0.0250	0.0101

We consider a simple location-scale DGP:

$$y_t = \mu + \sigma_\varepsilon \varepsilon_t \tag{22}$$

with:

$$\varepsilon_t \sim \mathsf{IN}[0, 1], \tag{23}$$

where $\mu = 0$ and $\sigma_\varepsilon = 1$. The aim is to investigate the impact on estimating μ and σ_ε^2 when saturating the model with impulse dummies.

We consider two econometric models. The first is given by:

$$y_t = \mu + \sum_{j=1}^{T-T/2} \delta_j d_{t,j} + \varepsilon_t \tag{24}$$

whilst the second is:

$$y_t = \mu + \sum_{t=T/2+1}^{T} \delta_j d_{t,j} + \varepsilon_t \tag{25}$$

T is the sample size and $d_{t,j}$ is a single impulse indicator. Hence, (24) contains $T/2$ impulse indicators for the first $T/2$ observations, and (25) contains $T/2$ impulse indicators for the last set of observations. Below, we consider alternative divisions of the indicators across the sample.

4.1 Empirical rejection frequencies of impulse indicators under the normal null

Given the DGP, the composite null hypothesis:

$$\mathsf{H}_0 : \delta_t = 0 \quad \forall\, t \tag{26}$$

is true, $\forall\, t$, for both models. We first estimate model (24) and then model (25), sequentially, under these assumptions, store the significant indicators, and combine these to obtain the final selected model and estimators (5) and (6). $M = 10,000$ replications were conducted for this experiment. From Hendry and Santos (2005), the OLS estimators of δ_t are unbiased with the usual Student $\mathsf{t}_{(T-T/2-1)}$ distributions under the null given by (26). Table 1 reports the mean rejection frequency of the null across the 10,000 experiments at a nominal significance *per test* of $\alpha = 0.05$, $\alpha = 0.025$ and $\alpha = 0.01$. A sample size of $T = 100$ is used. As expected, we obtain empirical rejection frequencies close to the nominals.

This outcome is not affected by randomly, rather than consecutively, adding $T/2$ dummies in each regression, unsurprisingly since the data have no time ordering.

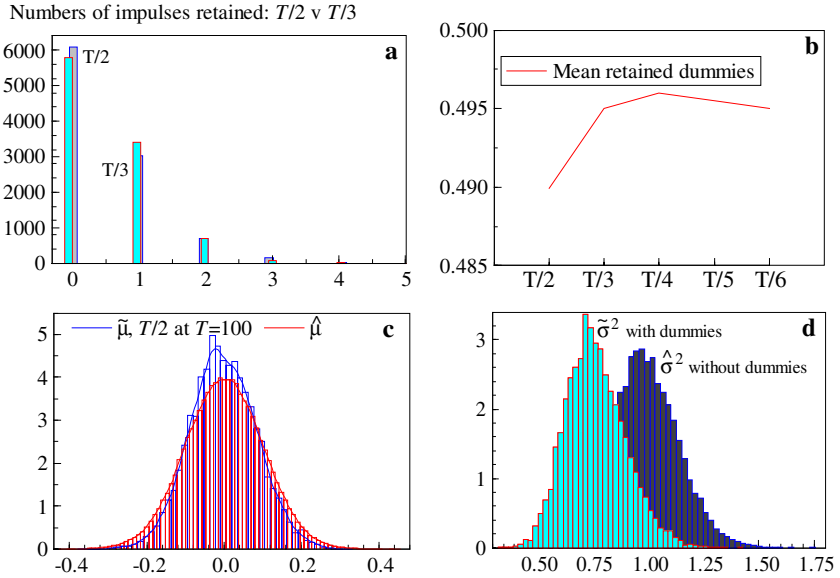

Fig. 1 Distribution of impulses, means and equations standard errors

Under an alternative where the break is a location shift, such shuffling could be useful (although this is the subject of a different paper).

4.1.1 Empirical distributions of retained impulses

Under the null hypothesis, the distributions of the numbers of empirically retained impulses are of interest: retention is decided on the basis of a two-sided individual significance test. We report these for $T = 50$, using the above settings, but including additional significance levels.

Figure 1a refers to $T = 50$ and uses a two-sided t-test with $\alpha = 0.01$. The x-axis measures the number of impulses retained *per* regression, and the y-axis the actual number of regressions (out of 10,000) that retained the given number of 'spurious' impulses.

The mode occurs at zero with probability $(1 - \alpha)^T \simeq 0.6$, with the probability of retaining one impulse by chance being $T\alpha \times (1 - \alpha)^{T-1} \simeq 0.3$. As Fig. 1a also shows, a three-way equal split of $T/3$ does not change the outcomes substantively: neither the mode nor the decay pattern alters. Corresponding outcomes would held at nominal significance levels of $\alpha = 0.025$ and $\alpha = 0.05$.

Figure 1b records the impact on the mean number of retained dummies of using finer equal sub-divisions of added impulses at $T = 50$ for $\alpha = 0.01$, so $\alpha T = 0.5$. There is very little variation in the number of retained dummies as the number of equal splits increases. The overall range of the mean estimate is 0.490–0.496, with the number of observations in the equal splits varying from $T/2$ to $T/6$.

Table 2 $\hat{\sigma}^2$ and $\tilde{\sigma}^2$: average across MC replications for $T = 50$ and $T = 100$, $\alpha = 0.01$

T	$\widehat{\sigma}^2$	$\widetilde{\sigma}^2$
50	0.977	0.901
100	0.989	0.910

4.2 Empirical distribution of $\widetilde{\mu}$ under the normal null

Figure 1c shows the empirical distributions of $\widetilde{\mu}$ and $\widehat{\mu}$ under the null for $T = 100$ and $\alpha = 0.01$. Throughout we shall use $\widehat{\mu}$ and $\widehat{\sigma}_\varepsilon^2$ as the full-sample OLS estimators of the mean and variance. $\widetilde{\mu}$ and $\widetilde{\sigma}_\varepsilon^2$ are the estimators for the impulse-saturated model. The distribution of $\widehat{\mu}$ is correctly centered, and more concentrated near the center, but as shown above, more dispersed in the tails, leading to a larger standard deviation.

4.3 Empirical distribution of $\widetilde{\sigma}_\varepsilon^2$ under the normal null

Figure 1d records the estimates of the residual variances for a sample size of $T = 100$, with ($\widetilde{\sigma}_\varepsilon^2$) and without ($\widehat{\sigma}_\varepsilon^2$) dummies, at $\alpha = 0.05$: the sampling distributions for $T = 50$ at the same settings were similar. As expected, $\widetilde{\sigma}_\varepsilon^2$ is downwards biased when impulses are introduced. Table 2 reports the average Monte Carlo estimates of σ_ε^2 at $\alpha = 0.01$. Since $\sigma_\varepsilon^2 = 1$, the expected downward biases in $\widetilde{\sigma}_\varepsilon^2$ are close to the value of $(1 - \frac{2c_\alpha \phi(c_\alpha)}{1-\alpha})\sigma_\varepsilon^2$ of -0.075243986. As the sample size increases, $\widetilde{\sigma}_\varepsilon^2$ is closer to the relevant limiting value.

4.4 Response surface for $\sigma_{\widetilde{\mu}}^2$ for normal errors

The distributional result in Sect. 3.1 was:

$$T^{1/2}(\widetilde{\mu} - \mu) \xrightarrow{\mathsf{D}} \mathsf{N}\left[0, \sigma_\varepsilon^2 \sigma_\mu^2\right], \tag{27}$$

so for normal errors when $\lambda_1 = \lambda_2$ from (8):

$$\sigma_\mu^2 = \frac{1}{(1-\alpha)}\left(1 + 4c_\alpha\phi(c_\alpha) - \frac{2c_\alpha\phi(c_\alpha)}{(1-\alpha)}\left[1 + 2c_\alpha\phi(c_\alpha)\right]\right)$$

and:

$$\left(\frac{T\mathsf{V}\left[\widetilde{\mu}\right]}{\sigma_\varepsilon^2}\right) = \sigma_\mu^2. \tag{28}$$

Thus, the simulations generated the values of the left-hand side of (28), which were then regressed on the numerical values of σ_μ^2 computed using (8).

The Monte Carlo simulations first confirmed the invariance of the outcomes from PcGets to the value of σ_ε^2 and to the form of 'split' into equal blocks of $m = 2$ and $m = 3$. There were 78 experiments spanning $c_\alpha = 5$ to $c_\alpha = 1$ ($\Phi(c_\alpha) \simeq 1$ to

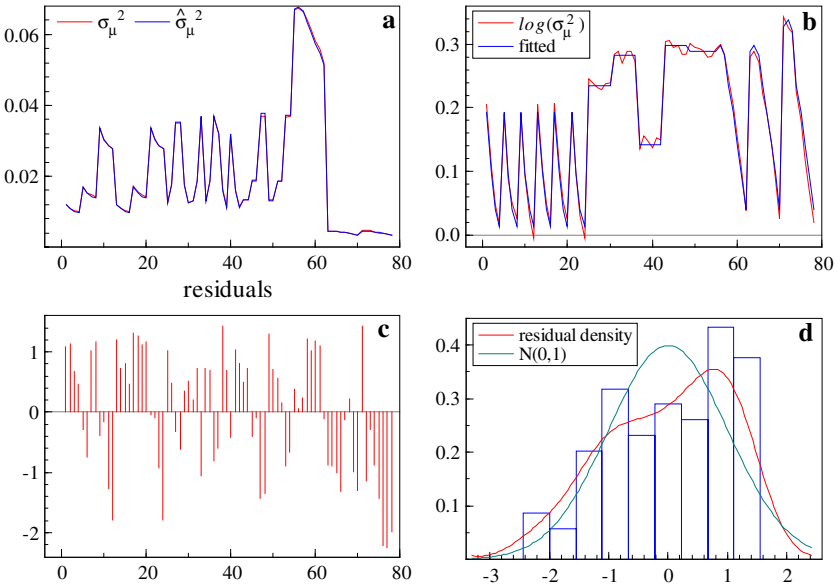

Fig. 2 Fitted versus actual values from the simulations, and residual analysis

$\Phi(c_\alpha) = 0.68$), and $T = 20$ to $T = 300$. The response surface for $\mathsf{V}[\widetilde{\mu}]$ yielded [heteroscedasticity consistent standard errors (HCSE) in parentheses: see White (1984)]:

$$\widehat{\mathsf{V}[\widetilde{\mu}]} = \underset{(0.0021)}{1.002}\, T^{-1}\sigma_\varepsilon^2\sigma_\mu^2, \tag{29}$$

$$\mathbf{R}^2 = 0.9997,\ \ \widehat{\sigma} = 1.4\%,\ \ \chi^2_{\mathsf{nd}}(2) = 16.4^{**},\ \ \mathsf{F}_{\mathsf{het}}(2,75) = 21.7^{**}. \tag{30}$$

Some outliers were detected and slightly alter the outcome, but as Fig. 2a shows, the fitted and actual values are extremely close across the 78 experiments. We also tested for whether the outcome depended on the split being in halves or in thirds and found that the corresponding dummy was insignificant.

The outcome using a scaled log form, reported here including the outlier correction for experiments 71–73, is given by:

$$\log\left(\frac{\widehat{T\mathsf{V}[\widetilde{\mu}]}}{\sigma_\varepsilon^2}\right) = \underset{(0.002)}{0.0135} + \underset{(0.011)}{0.936}\,\log\left(\sigma_\mu^2\right) + \underset{(0.006)}{0.04}\ I_{71-73}, \tag{31}$$

$$\mathbf{R}^2 = 0.9899,\quad \widehat{\sigma} = 1.04\%.$$

Figure 2b shows the fitted and actual values of (31) across the 78 experiments. The fit is again extremely close. Figure 2c, d plots, respectively, the estimation residuals, and the residuals estimated density.

Table 3 Summary results for a location-scale DGP with $\mathsf{t}_{(4)}$-distributed errors and $\mu = 0$; split at $T/2$

$T = 300$	$\left\|\mathsf{t}_{\hat{\delta}_t}\right\| > 2$	$\left\|\mathsf{t}_{\hat{\delta}_t}\right\| > 2.5$
$\mathsf{E}[\widetilde{\mu}]$	−0.002	−0.008
$\mathsf{V}[\widetilde{\mu}]$	0.00544	0.00535
ARNI	15.64	8.08
RF	0.052	0.027

4.5 Non-normality

We briefly consider the impact of saturation in a highly non-normal case, namely a $\mathsf{t}_{(4)}$-distributed error. Although this distribution does not satisfy the assumptions of Theorems 1 and 3, it was of interest to see if 'fat-tails' led to an excess of retained impulses.

A sample size of $T = 300$ was considered, for a sample split of $T/2$. At each replication, the T draws are from a $\mathsf{t}_{(4)}$ distribution. The moments of $X \sim \mathsf{t}_{(4)}$ are such that $\mathsf{E}[X] = 0$ and $\mathsf{V}[X] = v/(v-2) = 2$, where v denotes the degrees of freedom. Hence, when no impulses are added, $\mathsf{V}[\bar{X}] = 2/300 = 0.0067$ and $\sqrt{\mathsf{V}[\bar{X}]} = 0.082$.

We use a location-scale DGP, $\mu = 0$, with $\mathsf{t}_{(4)}$ errors. We consider two criteria for retention of single impulse indicators: $|\mathsf{t}_{\hat{\delta}_t}| > 2$ and $|\mathsf{t}_{\hat{\delta}_t}| > 2.5$.

Table 3 reports summary statistics from the Monte Carlo experiments, where ARNI stands for the average number of retained impulses in each replication. There is little evidence of an excess retention of impulses. The intuitive explanation is that the fat tails generate a much larger residual error variance, so only draws far into the tails are significant, even though nominal critical values relevant to the normal are used.

Conclusion

We have considered a problem that previously seemed intractable: selecting a regression when there are more regressors than observations. The special case we examined was for saturating the model with individual impulse indicators, one for each observation. A variant of the general-to-specific approach nevertheless suggested a feasible solution. Aspects of the distributions of the mean, its standard error, and the residual standard deviation, after retaining only significant impulses from the saturating set, were derived, together with an approximate operational bias correction for the last of these.

To select a regression when there are more regressors than observations requires both a block implementation of multi-path searches, as well as such procedures within tentative models as in PcGets. The Monte Carlo simulations based on doing so match the theoretical analysis, confirming that the approach is viable, with the null rejection frequencies as established above. Evidence suggests that the algorithm might also be of interest with fat-tailed distributions, that do not satisfy the assumptions of the main theorems. Indeed, even for the case of a $\mathsf{t}_{(4)}$ distribution, the average number of retained impulses is not excessive.

Moreover, many new problems become amenable to solution, including general regression sub-set selection, non-linear model selection, and new automatically computable tests of economic interest (see Hendry and Santos 2006).

Clearly, the task of selecting single impulse indicators in a saturated regression is made easier due to the nonexistence of collinearity problems that would arise with other types of regressors. The vectors of single impulse indicators are orthogonal to each others, which would not happen in general with other regressors.

References

Campos J, Hendry DF, Krolzig H-M (2003) Consistent model selection by an automatic gets approach. Oxf Bull Econ Stat 65:803–819

Campos J, Ericsson NR, Hendry DF (2005) Editor's introduction. In: Campos J, Ericsson NR, Hendry DF (eds) Readings on general-to-specific modelling. Edward Elgar, Cheltenham (forthcoming)

Doornik JA (2001) OX, an object oriented matrix programming language, 4th edn. Timberlake Consultants Press, London

Foster DP, Stine RA (2004) Honest confidence intervals for the error variance in stepwise regression, Mimeo. Statistics Department, Wharton School, University of Pennsylvania

Granger CWJ, Hendry DF (2003) A dialogue concerning a new instrument for econometric modelling. Unpublished Paper, Department of Economics, University of Oxford

Hendry DF (1995) Dynamic econometrics. Oxford University Press, Oxford

Hendry DF, Krolzig H-M (2001) Automatic econometric model selection using PcGets. Timberlake Consultants Press, London

Hendry DF, Krolzig H-M (2003) New developments in automatic general-to-specific modelling. In: Stigum BP (ed) Econometrics and the philosophy of economics. Princeton University Press, Princeton, pp 379–419

Hendry DF, Krolzig H-M (2004) Model selection with more variables than observations. Unpublished Paper, Department of Economics, University of Oxford

Hendry DF, Krolzig H-M (2005) The properties of automatic GETS modelling. Econ J 115:C32–C61

Hendry DF, Santos C (2005) Regression models with data-based indicator variables. Oxf Bull Econ Stat 67:571–595

Hendry DF, Santos C (2006) Automatic tests for super exogeneity. Unpublished Paper, Department of Economics, University of Oxford

Hoover KD, Perez SJ (1999) Data mining reconsidered: encompassing and the general-to-specific approach to specification search. Econometrics J 2:167–191

Hoover SJ, Perez SJ (2004) Truth and robustness in cross-country growth regressions. Oxf Bull Econ Stat 66(5):765–798

Krolzig H-M, Hendry DF (2001) Computer automation of general-to-specific model selection procedures. J Econ Dyn Control 25:831–866

Salkever DS (1976) The use of dummy variables to compute predictions, prediction errors, and confidence intervals. J Econometrics 4:393–397

White H (1984) Asymptotic theory for econometricians. Academic, London